基于分歧的半监督学习

周志华 [1]

摘 要 传统监督学习通常需使用大量有标记的数据样本作为训练例, 而在很多现实问题中, 人们虽能容易地获得大批数据样本, 但为数据提供标记却需耗费很多人力物力. 那么, 在仅有少量有标记数据时, 可否通过对大量未标记数据进行利用来提升学习性能呢? 为此, 半监督学习成为近十多年来机器学习的一大研究热点. 基于分歧的半监督学习是该领域的主流范型之一, 它通过使用多个学习器来对未标记数据进行利用, 而学习器间的 "分歧" 对学习成效至关重要. 本文将综述简介这方面的一些研究进展.

关键词 机器学习, 半监督学习, 基于分歧的半监督学习, 未标记数据

引用格式 周志华. 基于分歧的半监督学习. 自动化学报, 2013, **39**(11): 1871−1878

DOI 10.3724/SP.J.1004.2013.01871

Disagreement-based Semi-supervised Learning

ZHOU Zhi-Hua[1]

Abstract Traditional supervised learning generally requires a large amount of labeled data as training examples; in many real tasks, however, although it is usually easy to acquire a lot of data, it is often expensive to get the label information. Can we improve the learning performance with limited amount of labeled data by exploiting the large amount of unlabeled data? For this purpose, semi-supervised learning has become a hot topic of machine learning during the past ten years. One of the mainstream paradigms, the disagreement-based semi-supervised learning, trains multiple learners to exploit the unlabeled data, where the "disagreement" among the learners is crucial. This article briefly surveys some research advances of this paradigm.

Key words Machine learning, semi-supervised learning, disagreement-based semi-supervised learning, unlabeled data

Citation Zhou Zhi-Hua. Disagreement-based semi-supervised learning. *Acta Automatica Sinica*, 2013, **39**(11): 1871−1878

传统监督学习通过对大量有标记的 (Labeled) 训练例进行学习以建立模型用于预测未见示例的标记. 这里的 "标记 (Label)" 是指示例所对应的输出, 例如, 在分类任务中标记就是示例的类别, 而在回归任务中标记就是示例所对应的实值输出. 随着人类收集、存储数据能力的高度发展, 在很多实际任务中可以容易地获取大批未标记 (Unlabeled) 数据, 而对这些数据赋予标记则往往需要耗费大量的人力物力. 例如在进行计算机辅助医学影像分析时, 可以从医院获得大量医学影像, 但如果希望医学专家把影像中的病灶全都标识出来则是不现实的. "有标记数据少、未标记数据多" 这个现象在网上应用中更为明显. 例如, 在进行 Web 网页推荐时, 需请用户标记出感兴趣的网页, 但很少有用户愿意花很多时间来提供标记, 因此有标记的网页数据比较少, 但 Web 上存在着无数的网页, 它们都可作为未标记数据来使用. 显然, 如果只使用少量 "昂贵的" 有标记数据进行学习, 那么所训练出的学习系统可能很难具有强泛化能力; 而忽略了大量 "廉价的" 未标记数据, 则是对数据资源极大的浪费.

那么, 在仅有少量有标记数据时, 可否通过对大量未标记数据进行利用来提升学习性能呢? 这个问题不仅在理论上有重要意义, 还直接影响到机器学习技术在现实任务中所能发挥的效用, 因此受到了机器学习界的高度重视.

半监督学习 (Semi-supervised learning)[1−3] 试图让学习器自动地对大量未标记数据进行利用以辅助少量有标记数据进行学习. 它可进一步划分为纯半监督学习和直推学习 (Transductive learning), 主要区别是后者假定未标记数据就是待测数据, 即学习的目的就是在这些未标记数据上取得最佳泛化性能. 换言之, 纯半监督学习基于开放世界假设, 希望学得模型能适用于学习过程中未观察到的数据; 而

收稿日期 2013-07-05 录用日期 2013-08-28
Manuscript received July 5, 2013; accepted August 28, 2013
国家重点基础研究发展计划 (973 计划) (2010CB327903), 国家自然科学基金 (61073097) 资助
Supported by National Basic Research Program of China (973 Program) (2010CB327903) and National Natural Science Foundation of China (61073097)
庆祝《自动化学报》创刊 50 周年专刊约稿
Invited Articles for the Special Issue for the 50th Anniversary of Acta Automatica Sinica
1. 南京大学计算机软件新技术国家重点实验室 南京 210023
1. National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023

直推学习基于封闭世界假设, 只试图对学习过程中观察到的未标记数据进行预测.

目前的半监督学习方法大致有四种主流范型 (Paradigm), 即基于生成式模型的方法、半监督 SVM 方法、基于图的方法和基于分歧的方法. 基于生成式模型的方法假设所有数据均由相同的生成式模型产生, 借助模型参数将未标记数据与学习目标联系起来, 通常利用 EM 算法根据极大似然来估计模型参数; 半监督 SVM 方法通过调整 SVM 的超平面和未标记数据的标记指派, 使得 SVM 在所有训练数据 (包括有标记和未标记数据) 上最大化间隔 (Margin); 基于图的方法通过考虑数据样本间的联系, 将数据集映射为一个图, 然后在图上进行标记信息的 "传播"; 基于分歧的方法通过使用多个学习器来对未标记数据进行利用, 在学习过程中将未标记数据作为多学习器间信息交互的平台. 在这四种主流范型中, 基于生成式模型的方法出现较早[4−6], 而基于分歧的方法、半监督 SVM 方法和基于图的方法的代表性工作分别于 2008 年[7]、2009 年[8]、2013 年[9] 获得国际机器学习大会 "十年最佳论文奖", 由此可看出这些范型的影响力和重要性.

基于分歧的半监督学习的研究始于 Blum 和 Mitchell 关于协同训练 (Co-training) 的工作[7], 该类技术较少受到模型假设、损失函数非凸性和数据规模问题的影响, 学习方法简单有效、理论基础相对坚实、适用范围较为广泛. 后续研究揭示出, 多学习器间的 "分歧" 对此类学习的成效至关重要, 由此而被命名为基于分歧的半监督学习[10]. 关于这方面的研究已有综述文章[3, 11], 本文将在其基础上结合一些最新进展, 对基于分歧的半监督学习的学习方法、理论探讨、应用发展做综述简介.

1 学习方法

1.1 多视图学习

在某些应用任务中, 一个数据集可能包含多个属性集, 此时每个数据样本同时拥有多个特征向量描述; 这里的每个属性集即被称为数据的一个 "视图 (View)". 例如, 一幅图像可由图像本身的可视信息来描述, 也可由其关联的文字信息来描述, 这时可视信息所对应的属性集合就形成了关于图像的一个视图, 而文字信息对应的属性集合则形成了另一个视图; 再如, 一张网页可以由其本身包含的信息来描述, 也可以由其他网页指向它的超链接所包含的信息来描述, 两者对应的属性集合分别形成了关于网页的两个视图.

基于分歧的半监督学习的起源、也是最著名的代表性方法 — "协同训练法"[7], 即是针对多视图数据而提出的. 协同训练法要求数据具有两个充分冗余且满足条件独立性的视图, "充分 (Sufficient)" 是指每个视图都包含足够产生最优学习器的信息, 此时对其中任一视图来说, 另一个视图则是 "冗余 (Redundant)" 的; 同时, 对类别标记来说这两个视图条件独立. 协同训练法的学习过程非常简单: 首先分别在每个视图上利用有标记样本训练一个分类器, 然后, 每个分类器从未标记样本中挑选若干标记置信度 (即对样本赋予正确标记的置信度) 高的样本进行标记, 并把这些 "伪标记 (Pseudo-labeled)" 样本 (即其标记是由学习器给出的) 加入另一个分类器的训练集中, 以便对方利用这些新增的有标记样本进行更新. 这个 "互相学习、共同进步" 的过程不断迭代进行下去, 直到两个分类器都不再发生变化, 或达到预先设定的学习轮数为止.

协同训练法在多视图数据上实验效果很好, 并在理论上得到证明[7]: 当两个充分冗余视图确实满足条件独立性时, 通过协同训练可以利用未标记样本把弱分类器 (在二分类问题上就是精度略高于 50 % 的分类器) 的精度提升到任意高.

协同正则法 (Co-regularization)[12] 是受协同训练法的启发提出的基于正则化框架的方法, 它试图直接最小化有标记样本上的错误率和两个视图上未标记样本的标记不一致性. 与协同训练法不同的是, 协同正则法中不涉及对未标记样本赋予伪标记的过程. 该方法有多种算法实现[13−14], 并可在信息论框架下解释其工作原理[15].

协同 EM 法 (Co-EM)[16] 是一种容易想到的多视图半监督学习方法, 它通过在两个视图上联合进行生成式模型的 EM 参数估计来进行学习, 可视为基于生成式模型的半监督学习与基于分歧的半监督学习之间的过渡方法. 该方法后来被推广为可使用 SVM 分类器[17], 并可用于无监督聚类学习. 此外还有其他一些基于多视图的方法, 例如 Ando 和 Zhang[18] 先利用未标记数据学得两个视图上的有效属性, 然后再基于这些有效属性学得最终分类器.

上述方法大多假设充分冗余视图满足条件独立性. 然而遗憾的是, 视图的条件独立性假设通常并不成立; 例如即使对 Blum 和 Mitchell 在提出协同训练法时用来做 "动机例子" 的网页分类问题来说, 网页本身的信息和指向它的超链接信息通常高度相关, 绝非条件独立. 事实上, 视图的条件独立性是一个过强的假设, Zhou 等[19] 显示出, 若该假设成立, 甚至只需单个有标记样本就可有效地进行半监督学习; 因为两个条件独立的充分属性集包含了足以导出一个距离度量可信子空间的信息, 在该子空间中直接基于样本与有标记正例的距离计算就可完成分类.

更严峻的问题是, 虽然不少应用任务中的数据

具有多视图, 但它们未必是充分视图, 此时是否可学习、如何有效地学习, 在理论上仍有待探讨 (见第 2.2 节). 另一方面, 大多数现实任务只拥有一个属性集, 即单视图数据, 基于多视图的方法难以使用.

1.2 单视图学习

由于协同训练法取得了巨大成功, 对单视图数据, 研究者们首先考虑的是可否将单视图 "划分" 为多视图, 然后直接使用协同训练法等多视图方法. 这方面一个重要的尝试是 Nigam 和 Ghani 的工作[16], 他们通过实验研究发现, 在属性集非常大、包含大量冗余属性时 (例如在文本数据上, 如果将每个字作为一个属性, 则属性集非常大, 其中有大量的冗余属性可起到相似的描述作用), 随机地把属性集划分为多视图后, 协同训练法已可取得很好的效果; 然而, 在其他数据上, 视图划分的尝试都失败了. 此后, Du 等[20] 通过实验研究显示出, 在拥有足够的有标记样本时, 有可能找出合适的视图划分 (若单视图中确实包含充分冗余多视图信息), 然而, 半监督学习之所以有益, 正是因为有标记样本不足, 因此这一结果对单视图划分为多视图的前景给出了悲观的结论. 后来虽有一些理论工作可能对视图划分有所启发[21], 但目前仍未出现有效可行的方法.

Goldman 和 Zhou[22] 提出了一种可用于单视图数据的协同训练法变体, 通过使用两种不同的决策树算法在相同属性集上生成两个不同的分类器, 然后按协同训练法的方式来进行分类器增强. 由于该方法不基于多视图, 因此协同训练法原有的理论支撑均不适用, 造成的巨大障碍是在选择未标记样本时难以估计标记置信度, 以及在进行最终预测时, 若两个分类器预测结果不同, 难以确定该取信哪一个分类器. 为克服障碍, 该方法强制必须使用决策树算法, 因为决策树可把样本空间划分为若干个等价类 (每个叶结点对应了一个等价类), 而这些等价类的置信度可通过 10 折交叉验证法进行估计, 这样, 由于预测时样本必然会落到某个叶结点上, 该等价类的估计置信度就可用来代替标记置信度. 显然, 该方法对基分类器的约束限制了其适用范围, 而反复的 10 折交叉验证会导致很大的计算开销; 此外, 由于该方法严重依赖 10 折交叉验证法估计标记置信度, 使得其只适用于有标记样本很多的情况.

Zhou 和 Li[23] 提出了一种 "三体训练法 (Tri-training)", 该方法从单视图训练集中产生三个分类器, 然后利用这三个分类器以 "少数服从多数" 的形式来产生伪标记样本, 例如, 若两个分类器将某个未标记样本预测为正类, 而第三个分类器预测为反类, 则该样本被作为伪标记正样本提供给第三个分类器进行学习. 由此, 该方法避免了对标记置信度的显式估计. 然而, 某些情形下, 多数分类器的预测结果可能是错误的, 此时, 从少数分类器的角度看, 它收到的是有 "标记噪音" 的样本. 受 Goldman 和 Zhou[22] 的启发, 三体训练法基于 Angluin 和 Laird[24] 关于标记噪声学习的理论结果, 导出了 "少数服从多数" 所需满足的条件, 在学习中的每一轮, 只需判断该条件是否成立, 即可决定是否基于伪标记样本进行分类器更新; 直观上来说, 该结果表明在一定条件下积累的标记噪声可利用大量未标记数据进行补偿. 最终训练完成后, 三个分类器通过投票机制作为一个分类器集成进行使用. 值得指出的是, 最初的三个分类器必须强于弱学习器, 并且具有较大的分歧 (即差异); 这可以通过 Bootstrap 采样机制产生 (类似于集成学习[25] 中的 Bagging 方法). 该方法的另一个好处是, 可以容易地推广到三视图数据上. 由于三体训练法不需多视图、不对基学习器有特定要求, 算法实现简单、便于应用, 因此和协同训练法成为基于分歧的半监督学习方法中最常用的技术, 有时也被并称为 Co-tri-training[26].

Li 和 Zhou[27] 将三体训练法进行推广, 以使用更多的分类器. 他们提出了协同森林法 (Co-Forest)[27], 该方法以随机森林[28] 的方式产生多个分类器, 然后使用 "少数服从多数" 的形式来为少数学习器产生伪标记样本. 由于使用多个分类器, 分类器间的差异很难保持, 因此, 在算法实现上协同森林法使用了多种差异引入机制来减缓学习过程的 "早熟".

可以看出, 三体训练法和协同森林法同时利用了半监督学习和集成学习机制, 从而获得了学习性能的进一步提升. Hady 和 Schwenker[29] 进一步对利用多视图或多个不同学习器的思想进行拓展, 提出了利用分歧分类器集成进行半监督学习的 CoBC 框架. 值得指出的是, 虽然半监督学习和集成学习均致力于提升泛化性能, 但二者的研究方法论有较大差别[30], 因此长期处于独立发展状态. 上述工作显示出二者可相互联系, Zhou 进一步指出两者互有助益[30], 而基于分歧的半监督学习自然成为半监督学习与集成学习研究之间的桥梁.

早期的半监督学习研究聚焦在分类任务上, 虽然回归通常与分类并列为预测学习的两大任务, 但半监督回归一直缺乏报道, 其主要原因之一是半监督学习中常用的聚类假设在回归任务上不成立, 而回归任务中的标记置信度估计较为困难. Zhou 和 Li[31] 首先对半监督回归进行研究, 提出了协同回归法 (COREG). 该方法在算法实现上基于不同的距离度量与/或不同的 k 值产生不同的 k 近邻回归学习器, 然后基于预测一致性来挑选标记置信度高的样本赋予伪回归标记; 其基本思路是, 标记置信度

越高的样本越有可能被赋予真实的标记, 而具有真实标记的样本应能较一致地体现出回归的内在规律, 因此, 被回归学习器以高置信度标记的样本应是使该回归学习器与训练集更一致的样本. 该方法可容易地推广到多视图数据上.

1.3 半监督主动学习

在半监督学习之外, 利用未标记样本学习的另一大类技术是主动学习[32], 它假设学习器对环境有一定的控制能力, 可以 "主动地" 向学习器之外的某个 "神谕 (Oracle)" 进行查询来获得样本的真实标记. 在主动学习中, 学习器自行挑选出一些未标记样本并通过神谕查询获得其真实标记, 然后将这些有标记样本加入训练集进行常规的监督学习, 其技术难点在于如何使用尽可能少的查询来显著提升泛化性能. 显然, 若能将半监督学习与主动学习相结合, 将有助于更好地对未标记样本进行利用.

基于分歧的半监督学习使用了多学习器, 使得它与主动学习中的 "多学习器查询 (Query by committee)"[33] 有天然的亲和力, 可在几乎不增加计算开销的条件下实现半监督主动学习. 具体来说, 在每一轮学习中, 对未标记样本而言, 如果多数学习器的预测结果与少数学习器不一致, 那么标记置信度最高的样本就由多数学习器赋予伪标记之后提供给少数学习器学习, 这是半监督学习过程; 如果各学习器的预测结果不一致, 那么不一致性最高的样本就被选择出来进行查询以获取真实标记, 这是主动学习过程. Wang 和 Zhou[34] 证明, 这样的半监督主动学习方法比单纯主动学习的样本复杂度更低; 换言之, 它能用更少的样本达到相当甚至更强的泛化性能.

Zhou 等[35] 提出了一种基于分歧的半监督主动学习方法用于增强相关反馈 (Relevance feedback) 的性能. 相关反馈是多媒体信息检索中的一种常用技术, 以图像检索为例, 具体而言, 系统将检索结果提供给用户后, 若用户不满意, 则可选择一些图像并标示出其是否与用户查询相关, 然后系统根据这些信息再重新进行检索. Zhou 等基于不同的距离度量产生两个排序学习器, 每个学习器给图像赋予一个 $[-1, +1]$ 之间的得分, $+1$ 意味着它认为该图像与用户查询相关、且置信度最高, -1 意味着它认为该图像与用户查询无关、且置信度最高, 0 则意味着它对该图像难以判定. 两个学习器分别将自己最确信的相关图像和无关图像传递给对方, 然后利用伪标记图像进行更新; 这个半监督学习过程可进行多轮. 最后, 系统结合两个学习器的排序以获得总排序, 排在最前面的若干幅图像作为检索结果反馈给用户, 因为它们是最确信的相关图像; 排在中间的若干幅 (置信度接近为 0) 被放入反馈池 (Feedback pool), 供用户在下一轮相关反馈时进行标示, 因为这些图像或是两个学习器都不确信、或是两者的判断存在显著矛盾, 仅凭系统自身已难以解决. 与传统相关反馈相比, 该技术不再是让用户自行从检索结果中挑选图像给予反馈, 避免了 (缺乏经验的) 用户对系统已学得很好的图像给予无用反馈, 使得用户反馈可为系统性能的提升给予更多的帮助, 从而增强了系统与用户交互的效用.

2 理论探讨

2.1 充分多视图

基于分歧的半监督学习的早期理论探讨大都针对多视图学习、以协同训练法为标本进行分析. Blum 和 Mitchell[7] 在提出协同训练法时证明了一个有趣的定理: 如果数据拥有的两个充分冗余视图满足条件独立性, 那么若假设空间 $\mathcal{H}_2$ 是噪声模型下 PAC 可学习的, 则给定弱分类器 $h_1 \in \mathcal{H}_1$ 和未标记数据, 假设空间 $(\mathcal{H}_1, \mathcal{H}_2)$ 对协同训练法可学习. 该定理显示出, 协同训练法可通过利用未标记数据把弱分类器的性能提升到任意精度. 此后, Dasgupta 等[36] 进一步证明, 分类器间的分歧程度是协同训练法泛化错误率的上界.

然而, 上述结果过于乐观了, 因为在实际任务中, 视图的条件独立性假设通常不成立.

为了放松条件独立性假设, Abney[37] 定义了弱依赖性: 给定标记 y, 令 x_1 和 x_2 分别表示样本 x 在两个视图中对应的示例, 则分类器 f 和 g 的条件依赖性可表示为 $d_y = \frac{1}{2}\sum_{y',y''} P(g(x_2) = y''|f(x_1) = y', y) - P(g(x_2) = y''|y)$; 若 $d_y \leq \frac{p_2(q_1-p_1)}{2p_1q_1}$, 则称 f 和 g 满足弱依赖关系, 其中 $p_1 = \min_{y'} P(f(x_1) = y'|y)$, $p_2 = \min_{y'} P(g(x_2) = y'|y)$, $q_1 = 1 - p_1$. Abney 证明[37], 若两个视图上的分类器满足弱依赖性, 则分类器间的分歧程度仍是协同训练法泛化错误率的上界.

但是, 弱依赖性在实际任务中往往也难以满足. 令 $\mathcal{X}_i^+$ 表示视图 $\mathcal{X}_i$ 上的正样本集, Balcan 等[38] 定义了 α- 膨胀性 (Expansion):

$$P(\boldsymbol{S}_1 \oplus \boldsymbol{S}_2) \geq \alpha \min[P(\boldsymbol{S}_1 \wedge \boldsymbol{S}_2), P(\boldsymbol{S}_1 \wedge \boldsymbol{S}_2)]$$

其中, $\boldsymbol{S}_i \subseteq \mathcal{X}_i^+$, $P(\boldsymbol{S}_1 \wedge \boldsymbol{S}_2)$ 表示被两个视图同时正确标记的正样本出现的概率, $P(\boldsymbol{S}_1 \oplus \boldsymbol{S}_2)$ 表示只被一个视图正确标记的正样本出现的概率, $\rho \leq P(\boldsymbol{S}_1^0 \cup \boldsymbol{S}_2^0)$ 表示正样本在至少一个视图上被初始分类器正确标记的概率下界. Balcan 等证明[38]: 若视图满足 α- 膨胀性, 并且每个视图上的分类器均能可信正确地标记正样本, 则协同训练法迭代 $O(\frac{1}{\alpha}\log\frac{1}{\epsilon} +$

$\frac{1}{\alpha\rho}$) 轮后可将初始分类器的错误率降低至 ϵ.

Balcan 等的 α- 膨胀性假设对视图假设做了进一步放松, 从而在一定程度上为实际任务中视图条件独立性假设、甚至弱依赖性假设都不成立时, 基于分歧的多视图半监督学习方法仍可能奏效的原因提供了解释. 需要注意的是, 他们对学习器的能力进行了约束, 要求每个视图上的分类器均能可信正确地标记正样本; 这为实际任务中使用强基学习器的惯例提供了理论依据.

2.2 充分单视图

上述理论结果均假设数据包含多个视图, 难以为单视图学习方法提供理论支撑.

Wang 和 Zhou[39] 证明, 只需两个 PAC 学习器具有较大的差异, 就可通过协同训练法利用未标记数据提升学习性能. 这一结果揭示出此类方法的本质是在利用学习器间的分歧, 而多视图只是为学习器产生分歧提供了更有利的条件; 若能通过其他机制为学习器产生足够的差异, 则在单视图条件下也可进行基于分歧的半监督学习, 从而为单视图学习方法提供了理论支撑. 值得一提的是, 以往理论分析显示出可通过利用未标记样本将学习器精度提升到任意高, 但实验和应用却显示出在运行一定的轮数后会出现 “饱和” 现象, 即进一步利用未标记样本不起作用; 换言之, 理论结果与实际效果之间存在一个大间隙. Wang 和 Zhou 的结果弥补了这个间隙: 由于学习器相互学习, 它们必然变得逐渐相似, 从而导致在运行一定的轮数后, 分类器间的分歧将不足以支持进一步的性能提高. 这为设计出 “自适应停止” 方法提供了启示.

进一步, Wang 和 Zhou[21] 将基于分歧的半监督学习方法转化为两个假设空间上的标记联合传播过程, 从而建立了半监督学习的两大主流范型, 即基于分歧的方法与基于图的方法之间的联系. 具体来说, 协同训练法基于有标记样本 x^s 和未标记样本 x^t 进行分类可看作估计后验概率 $P\big(y(x^t) = y(x^s)|x^t, x^s\big)$, 该后验概率矩阵可规范化为概率转移矩阵, 从而对应了一个图传播结构. 在每个假设空间对应的图 P_i 上, 有标记样本的标记被传播给一些未标记样本, 这些伪标记样本加入另一个图, 并通过此图传播给更多的未标记样本; 这样, 学习的成效直接与图 P_i 的性质有关. 由此, Wang 和 Zhou[21] 证明了协同训练法的充分必要性定理. 该结果显示出, 协同训练只关心权值矩阵的性质, 而并不在意权值矩阵是否通过多视图得到, 这确认了基于分歧的学习方法并不需要多视图, 仅要求分类器间存在适当的分歧; 而必要性条件是每个未标记样本在联合图中都与有标记样本连通. 这是协同训练法诞生 12 年来首次得到关于其奏效条件的充分必要性定理, 从而使其具备了相对完整的理论刻画. 因此该工作发布后即受到重视, 例如 Tom Mitchell 在国际机器学习大会上的 Keynote 报告及一系列 Distinguished lectures 中介绍了此结果. 另一方面, 以往有不少利用多个权值矩阵或拉普拉斯矩阵相结合来进行分类的方法[40−42], 但奏效的原因一直不清楚, Wang 和 Zhou 的这一结果在某种程度上为其提供了理论支撑.

2.3 不充分视图

上述理论探讨都基于一个共同的假设: 视图是充分的, 即视图可提供足够的信息来正确预测所有样本的标记. 在多视图情形下, 这意味着每个视图都可提供足够信息以学得能将所有样本正确分类的完美分类器. 基于这一假设, Balcan 和 Blum[43] 利用相容性 (Compatibility) 的概念提出了一种关于多视图半监督学习的 PAC 框架. 值得注意的是, 如果不能保证每个视图提供足够信息来正确预测所有样本的标记, 则每个视图上的最优分类器都会错误地标记某些样本, 这会导致不同视图上的最优分类器不相容. 因此, Balcan 和 Blum 的相容性 PAC 框架不适用于不充分视图上的学习.

在现实任务中, 由于属性退化和各种噪声的存在, 几乎无法保证视图信息的充分性, 因此, 探讨基于分歧的半监督学习在不充分视图上何时可行, 是一个非常基础而重要的问题. 然而, 这方面的研究具有高度的挑战性, 目前很少有报道. 直觉上来说, 在不充分视图上, 若视图之间具有较好的互补性, 则学习仍有可能奏效. 因此, 研究者们先针对不充分多视图进行探讨.

Wang 和 Zhou[44] 在这方面得到了一些结果, 他们的分析指出: 在不充分视图上, 学习过程会受到 “标记噪声” 和 “采样偏差” 的制约, 仅以协同训练法的方式通过学习器相互提供伪标记样本很难学得近似最优分类器. 但如果基学习器在提供预测结果之外, 还可提供对预测结果置信度的估计, 例如分类间隔 (Margin), 则可通过 “自适应” 提供不同数量的伪标记样本 (例如在间隔差较大时令学习器提供较多的伪标记样本, 而在间隔差较小时提供较少的伪标记样本), 在一定程度上缓解标记噪声和采样偏差的制约, 从而利用未标记样本提升学习性能. 这一结果显示出, 在不充分视图上, 基于分歧的半监督学习在某些条件下仍是可行的.

3 应用发展

基于分歧的半监督学习技术已在众多领域中得到广泛应用, 本节仅撷几例说明.

在自然语言处理领域, 研究者们甚至在协同训练法提出之前就意识到可利用任务本身涉及的不同属性集来建立更好的模型. 例如, Yarowsky[45] 通过同时使用词的局部上下文及词在文档其他部分出现时的含义进行词义消歧; Riloff 和 Jones[46] 同时考虑名词短语本身及其出现的上下文对名词短语进行地理位置分类; Collins 和 Singer[47] 同时使用名实体的拼写信息及名实体出现的上下文信息进行名实体识别. 而在协同训练法出现之后, 基于分歧的半监督学习技术受到很大重视, 不仅出现了很多实际应用, 还派生出不少融合了自然语言处理特点的变体算法[11].

在多媒体信息处理领域, Zhou 等[35] 提出基于分歧的半监督主动学习方法用于增强图像检索中相关反馈机制的性能, 该技术还可在视频检索等方面加以应用. Bai 等[48] 提出融合多种相似度度量的鲁棒形状检索半监督学习框架, 在 MPEG-7 数据集上的结果显著优于经典方法. Wang 等[49] 对三体训练法进行拓展, 提出 Tri-tracking 物体跟踪框架.

基于分歧的半监督学习技术最近还被成功应用于微处理器设计领域. 微处理器设计的一个重要环节是设计空间探索 (Design space exploration, DSE), 即选择适当的设计参数配置来满足需求. 设计空间的规模随设计参数的增加而指数级增长, 而常规 DSE 技术对每个配置进行评估, 需付出相当高的仿真代价. Guo 等[50] 将基于分歧的半监督学习技术引入 DSE, 仅对少量设计配置进行仿真并作为有标记样本, 而将其他设计配置作为未标记样本; 考虑到设计需求同时涉及性能和功耗, 他们使用了两个模型树学习器来利用未标记数据进行学习; 最终根据预测结果选择 "好的" 设计配置进行仿真评估. 在 SPEC CPU2000 基准测试上, 该技术获得了比现有技术 30 %∼84 % 的精度提升, 从而大幅降低了仿真成本, 并在龙芯 3B 的设计中得以应用.

此外, 在网络信息处理领域, 三体训练法被应用于垃圾邮件检测[51]、P2P (Peer-to-peer) 网络流量分类[52] 等, Tang 和 Han[53] 还进行了推广, 使其可以更好地应用于增量式学习环境.

4 展望

在一定程度上可以说, 过去十五年是半监督学习取得大发展的 "黄金十五年", 而 1998 年协同训练法这一基于分歧的半监督学习范型代表性方法的提出, 直接掀起了半监督学习研究的热潮. 基于分歧的半监督学习技术在不少方面很有吸引力, 例如它可视为一种 "元学习" 方法, 只需采用合适的基学习器, 就较少受到模型假设、损失函数非凸性和数据规模问题的影响, 学习方法简单有效、理论基础相对坚实、适用范围较为广泛, 并且为半监督学习与集成学习这两个基于不同方法论提升泛化性能的机器学习分支提供了天然桥梁. 本文简介了基于分歧的半监督学习研究进展, 更多信息可参见文献 [3, 11].

经过多年发展, 半监督学习技术在很多方面已趋于成熟, 并逐渐在实际应用中发挥作用. 然而, 该领域仍有一些重要问题有待研究. 笔者认为, 目前最重要的圣杯问题是 "安全" 半监督学习问题, 即利用未标记数据后期望性能会有所提升、同时还必须确信性能不会坏于只利用有标记数据学习.

半监督学习的倡导者以往告诉人们, 通过利用未标记数据可带来学习性能的显著提升. 然而, 很多研究发现, 在不少情况下, 利用未标记数据之后却可能带来性能 "恶化", 即泛化性能甚至不如只利用有标记数据学得的模型. 这一现象早有报道[54], 其成因在基于生成式模型的方法上相对清楚: 若模型假设不正确, 则对未标记数据利用越多, 性能越坏; 因此可试图依赖更丰富、可靠的领域知识来设计模型, 从而降低生成式方法性能恶化的风险. 但是, 对生成式方法之外的更通用的半监督学习范型, 安全半监督学习一直是未决的难题. Li 和 Zhou[55] 曾尝试使用数据审计 (Data editing) 技术来发现并去除 "坏的" 伪标记样本, 但由于数据审计技术严重依赖邻域信息, 只在数据密度高时才有帮助. 最近, Li 和 Zhou[56] 在安全半监督 SVM 学习上取得进展, 提出了 S4VM 方法. 但对基于分歧的半监督学习技术, 目前仍缺乏有效的解决方案.

另一方面, 半监督学习之所以引人注目, 重要缘由是随着机器学习技术走向实际应用, 人们发现, 在众多的现实任务中可以廉价地获得大批未标记数据, 但获取标记信息却相当昂贵, 因为对样本的标记过程通常需依赖专家知识或消耗物质资源; 因此, 有效地利用未标记数据成为关注的焦点. 然而, 在目前 Crowd sourcing 大发展的年代, 情况有所改观, 因为通过 Crowd sourcing, 可以用很小的代价获得大量标记信息, 例如人们每天在网上输入 "验证码", 为 OCR 任务提供了标记信息; 在网上书店、游戏网站表达倾向、发表评论, 也为相关商品提供了标记信息. 此类标记信息质量较差, 不仅存在大量缺失、噪音, 还可能存在恶意误导; 但无论如何, 与未标记数据相比, 信息量仍大了许多. 笔者预见, 有效利用低质量标记数据的技术, 在今后的十多年中可望取得大发展; 而基于分歧的半监督学习不仅将在标记信息难以通过 Crowd sourcing 轻易获得的任务中继续发挥重要作用 (例如医学影像分析中涉及的标记信息高度依赖于医学专家的专业知识, 很难通过廉价的 Crowd sourcing 获得), 其中的一些思想还对低质量标记数据的利用有可供借鉴之处.

References

1 Chapelle O, Scholkopf B, Zien A. *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006

2 Zhu X J. Semi-supervised Learning Literature Survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006

3 Zhou Z H, Li M. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 2010, **24**(3): 415−439

4 Shahshahani B M, Landgrebe D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 1994, **32**(5): 1087−1095

5 Miller D, Uyar H. A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, 1997. 571−577

6 Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000, **39**(2−3): 103−134

7 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. New York, USA: ACM, 1998. 92−100

8 Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. 200−209

9 Zhu X J, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning. Menlo Park, CA: AAAI Press, 2003. 912−919

10 Zhou Z H. Semi-supervised learning by disagreement. In: Proceedings of the 4th IEEE International Conference on Granular Computing. Piscataway, NJ: IEEE, 2008. 93

11 Zhou Zhi-Hua. Co-training paradigm of semi-supervised learning. *Machine Learning and Applications*. Beijing: Tsinghua University Press, 2007. 259−275
(周志华. 半监督学习中的协同训练风范. 机器学习及其应用. 北京: 清华大学出版社, 2007. 259−275)

12 Sindhwani V, Niyogi P, Belkin M. A co-regularized approach to semi-supervised learning with multiple views. In: Proceedings of the 22nd International Conference on Machine Learning. Cambridge, MA: MIT Press, 2005. 824−831

13 Brefeld U, Gartner T, Scheffer T, Wrobel S. Efficient co-regularised least squares regression. In: Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM, 2006. 137−144

14 Farquhar J D R, Hardoon D R, Meng H Y, Shawe-Taylor J, Szedmak S. Two view learning: SVM-2K, theory and practice. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2006. 355−362

15 Sridharan K, Kakade S M. An information theoretic framework for multi-view learning. In: Proceedings of the 21st Annual Conference on Learning Theory. Berlin, Germany: Springer, 2008. 403−414

16 Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In: Proceedings of the 9th International Conference on Information and Knowledge Management. New York, USA: ACM, 2000. 86−93

17 Brefeld U, Scheffer T. Co-EM support vector learning. In: Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM, 2004. 16−23

18 Ando R K, Zhang T. Two-view feature generation model for semi-supervised learning. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 25−32

19 Zhou Z H, Zhan D C, Yang Q. Semi-supervised learning with very few labeled training examples. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2007. 675−680

20 Du J, Ling C X, Zhou Z H. When does co-training work in real data? *IEEE Transactions on Knowledge and Data Engineering*, 2010, **23**(5): 788−799

21 Wang W, Zhou Z H. A new analysis of co-training. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: ICML, 2010. 1135−1142

22 Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data. In: Proceedings of the 17th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2000. 327−334

23 Zhou Z H, Li M. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17**(11): 1529−1541

24 Angluin D, Laird P. Learning from noisy examples. *Machine Learning*, 1988, **2**(4): 343−370

25 Zhou Z H. *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: Chapman and Hall/CRC, 2012

26 Breve F, Zhao L, Quiles M, Pedrycz W, Liu J M. Particle competition and cooperation in networks for semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2012, **24**(9): 1686−1698

27 Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 2007, **37**(6): 1088−1098

28 Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5−32

29 Hady M F A, Schwenker F. Co-training by committee: a generalized framework for semi-supervised learning with committees. *International Journal of Software and Informatics*, 2008, **2**(2): 95−124

30 Zhou Z H. When semi-supervised learning meets ensemble learning. In: Proceedings of the 8th International Workshop on Multiple Classifier Systems. Berlin, Germany: Springer, 2009. 529−538

31 Zhou Z H, Li M. Semi-supervised regression with co-training. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2005. 908−913

32 Settles B. Active Learning Literature Survey. Technical Report 1648, Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI, 2009

33 Seung H S, Opper M, Sompolinsky H. Query by committee. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. New York, USA: ACM, 1992. 287−294

34 Wang W, Zhou Z H. On multi-view active learning and the combination with semi-supervised learning. In: Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008. 1152−1159

35 Zhou Z H, Chen K J, Dai H B. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 2006, **24**(2): 219−244

36 Dasgupta S, Littman M L, McAllester D. PAC generalization bounds for co-training. *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2001. 375−382

37 Abney S. Bootstrapping. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2002. 360−367

38 Balcan M F, Blum A, Yang K. Co-training and expansion: towards bridging theory and practice. *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press, 2005. 89−96

39 Wang W, Zhou Z H. Analyzing co-training style algorithms. In: Proceedings of the 18th European Conference on Machine Learning. Berlin, Heidelberg: Springer-Verlag, 2007. 454−465

40 Argyriou A, Herbster M, Pontil M. Combining graph Laplacians for semi-supervised learning. *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2006. 67−74

41 Zhang T, Popescul A, Dom B. Linear prediction models with graph regularization for web-page categorization. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2006. 821−826

42 Zhou D Y, Burges C J C. Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 1159−1166

43 Balcan M F, Blum A. A discriminative model for semi-supervised learning. *Journal of the ACM*, 2010, **57**(3): Article 19

44 Wang W, Zhou Z H. Co-training with insufficient views. In: Proceedings of the 5th Asian Conference on Machine Learning. Canberra, Australia: ACML, 2013

45 Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 1995. 189−196

46 Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping. In: Proceedings of the 16th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1999. 474−479

47 Collins M, Singer Y. Unsupervised models for named entity classification. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. New Brunswick, NJ: ACL, 1999. 100−110

48 Bai X, Wang B, Yao C, Liu W Y, Tu Z W. Co-transduction for shape retrieval. *IEEE Transactions on Image Processing*, 2012, **21**(5): 2747−2757

49 Wang D, Yang G, Lu H C. Tri-tracking: combining three independent views for robust visual tracking. *International Journal of Image and Graphics*, 2012, **12**(3): 1250021

50 Guo Q, Chen T S, Chen Y J, Zhou Z H, Hu W W, Xu Z W. Effective and efficient microprocessor design space exploration using unlabeled design configurations. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2011. 1671−1677

51 Mavroeidis D, Chaidos K, Pirillos S, Christopoulos D, Vazirgiannis M. Using tri-training and support vector machines for addressing the ECML-PKDD 2006 discovery challenge. In: Proceedings of the 2006 ECML-PKDD Discovery Challenge Workshop. Berlin, Germany: ECML-PKDD, 2006. 39−47

52 Raahemi B, Zhong W C, Liu J. Exploiting unlabeled data to improve peer-to-peer traffic classification using incremental tri-training method. *Peer-to-Peer Networking and Applications*, 2009, **2**(2): 87−97

53 Tang X L, Han M. Ternary reversible extreme learning machines: the incremental tri-training method for semi-supervised classification. *Knowledge and Information Systems*, 2010, **23**(3): 345−372

54 Cozman F G, Cohen I. Unlabeled data can degrade classification performance of generative classifiers. In: Proceedings of the 15th International Conference of the Florida Artificial Intelligence Research Society. Pensacola, FL: AAAI Press, 2002. 327−331

55 Li M, Zhou Z H. SETRED: self-training with editing. In: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer-Verlag, 2005. 611−621

56 Li Y F, Zhou Z H. Towards making unlabeled data never hurt. In: Proceedings of the 28th International Conference on Machine Learning. Bellevue, WA: ICML, 2011. 1081−1088

周志华 博士, 南京大学计算机科学与技术系教授, 教育部长江学者特聘教授, IEEE Fellow, IAPR Fellow. 主要研究方向为人工智能, 机器学习, 数据挖掘, 模式识别, 多媒体信息检索.
E-mail: zhouzh@nju.edu.cn
(**ZHOU Zhi-Hua** Ph. D., Cheung Kong professor in the Department of Computer Science and Technology, Nanjing University. IEEE Fellow, IAPR Fellow. His research interest covers artificial intelligence, machine learning, data mining, pattern recognition, and multimedia information retrieval.)